Monthly Notices
of the
ROYAL ASTRONOMICAL SOCIETY

MNRAS **478,** 5547–5563 (2018)
Advance Access publication 2018 May 18

doi:10.1093/mnras/sty1308

Radio Galaxy Zoo: machine learning for radio source host galaxy cross-identification

M. J. Alger,[1,2]★ J. K. Banfield,[1,3] C. S. Ong,[2,4] L. Rudnick,[5] O. I. Wong,[3,6] C. Wolf,[1,3] H. Andernach,[7] R. P. Norris[8,9] and S. S. Shabala[10]

[1]*Research School of Astronomy and Astrophysics, The Australian National University, Canberra, ACT 2611, Australia*
[2]*Data61, CSIRO, Canberra, ACT 2601, Australia*
[3]*ARC Centre of Excellence for All-Sky Astrophysics (CAASTRO)*
[4]*Research School of Computer Science, The Australian National University, Canberra, ACT 2601, Australia*
[5]*Minnesota Institute for Astrophysics, University of Minnesota, 116 Church St. SE, Minneapolis, MN 55455, USA*
[6]*International Centre for Radio Astronomy Research-M468, The University of Western Australia, 35 Stirling Hwy, Crawley, WA 6009, Australia*
[7]*Departamento de Astronomía, DCNE, Universidad de Guanajuato, Apdo. Postal 144, CP 36000, Guanajuato, Gto., Mexico*
[8]*Western Sydney University, Locked Bag 1797, Penrith South, NSW 1797, Australia*
[9]*CSIRO Astronomy and Space Science, PO Box 76, Epping, NSW 1710, Australia*
[10]*School of Natural Sciences, University of Tasmania, Private Bag 37, Hobart, Tasmania 7001, Australia*

Accepted 2018 May 15. Received 2018 May 14; in original form 2017 November 2

ABSTRACT

We consider the problem of determining the host galaxies of radio sources by cross-identification. This has traditionally been done manually, which will be intractable for wide-area radio surveys like the Evolutionary Map of the Universe. Automated cross-identification will be critical for these future surveys, and machine learning may provide the tools to develop such methods. We apply a standard approach from computer vision to cross-identification, introducing one possible way of automating this problem, and explore the pros and cons of this approach. We apply our method to the 1.4 GHz Australian Telescope Large Area Survey (ATLAS) observations of the Chandra Deep Field South (CDFS) and the ESO Large Area ISO Survey South 1 fields by cross-identifying them with the Spitzer Wide-area Infrared Extragalactic survey. We train our method with two sets of data: expert cross-identifications of CDFS from the initial ATLAS data release and crowdsourced cross-identifications of CDFS from Radio Galaxy Zoo. We found that a simple strategy of cross-identifying a radio component with the nearest galaxy performs comparably to our more complex methods, though our estimated best-case performance is near 100 per cent. ATLAS contains 87 complex radio sources that have been cross-identified by experts, so there are not enough complex examples to learn how to cross-identify them accurately. Much larger data sets are therefore required for training methods like ours. We also show that training our method on Radio Galaxy Zoo cross-identifications gives comparable results to training on expert cross-identifications, demonstrating the value of crowdsourced training data.

Key words: methods: statistical – techniques: miscellaneous – galaxies: active – infrared: galaxies – radio continuum: galaxies.

1 INTRODUCTION

Next generation radio telescopes such as the Australian SKA Pathfinder (ASKAP; Johnston et al. 2007) and Apertif (Verheijen et al. 2008) will conduct increasingly wide, deep, and high-resolution radio surveys, producing large amounts of data. The Evolutionary Map of the Universe (EMU; Norris et al. 2011) survey using ASKAP is expected to detect over 70 million radio sources, compared to the 2.5 million radio sources currently known (Banfield et al. 2015). An important part of processing these data is cross-identifying observed radio emission regions with observations of their host galaxy in surveys at other wavelengths.

In the presence of extended radio emission cross-identification of the host can be a difficult task. Radio emission may extend far from the host galaxy and emission regions from a single physical object may appear disconnected. As a result, the observed structure of a radio source may have a complex relationship with the

★ E-mail: matthew.alger@anu.edu.au

corresponding host galaxy, and cross-identification in radio is much more difficult than cross-identification at shorter wavelengths. Small surveys containing a few thousand sources such as the Australia Telescope Large Area Survey (ATLAS; Norris et al. 2006; Middelberg et al. 2008) can be cross-identified manually, but this is impractical for larger surveys.

One approach to cross-identification of large numbers of sources is crowdsourcing, where volunteers cross-identify radio sources with their host galaxy. This is the premise of Radio Galaxy Zoo[1] (RGZ, Banfield et al. 2015), a citizen science project hosted on the Zooniverse platform (Lintott et al. 2008). Volunteers are shown radio and infrared images and are asked to cross-identify radio sources with the corresponding infrared host galaxies. An explanation of the project can be found in Banfield et al. (2015). The first data release for RGZ will provide a large data set of over 75 000 radio-host cross-identifications and radio source morphologies (Wong et al., in preparation). While this is a much larger number of visual cross-identifications than have been made by experts (e.g. Norris et al. 2006; Taylor et al. 2007; Middelberg et al. 2008; Gendre & Wall 2008; Grant et al. 2010), it is still far short of the millions of radio sources expected to be detected in upcoming radio surveys (Norris 2017a).

Automated algorithms have been developed for cross-identification. Fan et al. (2015) applied Bayesian hypothesis testing to this problem, fitting a three-component model to extended radio sources. This was achieved under the assumption that extended radio sources are composed of a core radio and two lobe components. The core radio component is coincident with the host galaxy, so cross-identification amounts to finding the galaxy coincident with the core radio component in the most likely model fit. This method is easily extended to use other, more complex models, but it is purely geometric. It does not incorporate other information such as the physical properties of the potential host galaxy. Additionally, there may be new classes of radio source detected in future surveys like EMU which do not fit the model. Weston et al. (2018) developed a modification of the likelihood ratio method of cross-identification (Richter 1975) for application to ATLAS and EMU. This method does well on non-extended radio sources with approximately 70 per cent accuracy in the ATLAS fields, but does not currently handle more complex (extended or multicomponent) radio sources (Norris 2017b).

One possibility is that machine learning techniques can be developed to automatically cross-identify catalogues drawn from new surveys. Machine learning describes a class of methods that learn approximations to functions. If cross-identification can be cast as a function approximation problem, then machine learning will allow data sets such as RGZ to be generalized to work on new data. Data sets from citizen scientists have already been used to train machine learning methods. Some astronomical examples can be found in Marshall, Lintott & Fletcher (2015).

In this paper, we cast cross-identification as a function approximation problem by applying an approach from computer vision literature. This approach casts cross-identification as the standard machine learning problem of binary classification by asking whether a given infrared source is the host galaxy or not. We train our methods on expert cross-identifications and volunteer cross-identifications from RGZ. In Section 2, we describe the data we use to train our methods. In Section 3, we discuss how we cast the radio host galaxy cross-identification problem as a machine learning problem. In Section 4, we present results of applying our method to ATLAS observations of the Chandra Deep Field South (CDFS) and the ESO Large Area ISO Survey South 1 (ELAIS-S1) field. Our data, code, and results are available at https://radiogalaxyzoo.github.io/atlas-xid.

Throughout this paper, a 'radio source' refers to all radio emission observed associated with a single host galaxy, and a 'radio component' refers to a single, contiguous region of radio emission. Multiple components may arise from a single source. A 'compact' source is composed of a single unresolved component. Equation (1) shows the definition of a resolved component. We assume that all unresolved components are compact sources, i.e. we assume that each unresolved component has its own host galaxy.[2] An 'extended' source is a non-compact source, i.e. resolved single-component sources or a multicomponent source. Fig. 1 illustrates these definitions.

2 DATA

We use radio data from the ATLAS (Norris et al. 2006; Franzen et al. 2015), infrared data from the Spitzer Wide-area Infrared Extragalactic survey (SWIRE; Lonsdale et al. 2003; Surace et al. 2005), and cross-identifications of these surveys from the citizen science project RGZ (Banfield et al. 2015). RGZ also includes cross-identifications of sources in Faint Images of the Radio Sky at Twenty-Centimeters (FIRST; White et al. 1997) and the All*WISE* survey (Cutri et al. 2013), though we focus only on RGZ data from ATLAS and SWIRE.

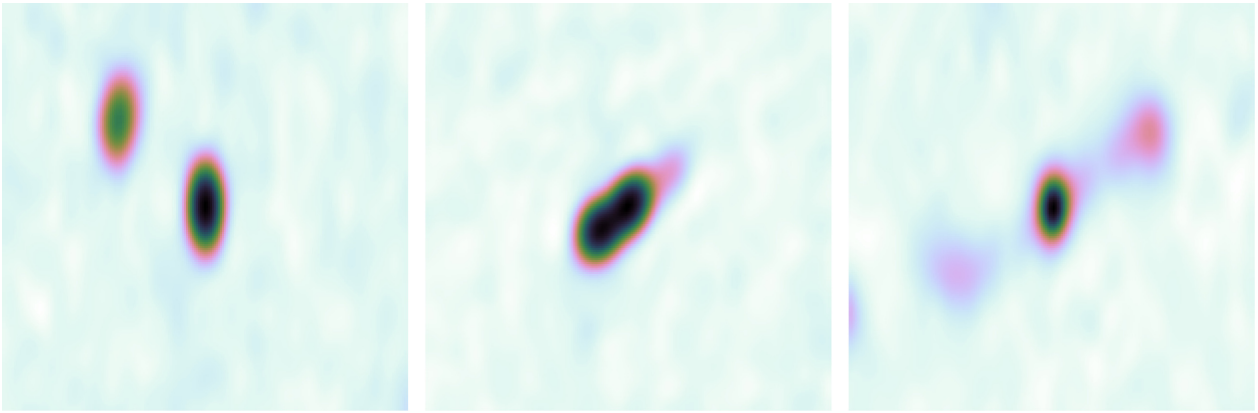
2.1 ATLAS

ATLAS is a pilot survey for the EMU (Norris et al. 2011) survey, which will cover the entire sky south of +30 deg and is expected to detect approximately 70 million new radio sources. 95 per cent of these sources will be single-component sources, but the remaining 5 per cent pose a considerable challenge to current automated cross-identification methods (Norris et al. 2011). EMU will be conducted at the same depth and resolution as ATLAS, so methods developed for processing ATLAS data are expected to work for EMU. ATLAS is a wide-area radio survey of the CDFS and ELAIS-S1 fields at 1.4 GHz with a sensitivity of 14 and 17 μJy beam^{-1} on CDFS and ELAIS-S1 respectively. CDFS covers 3.6 deg^2 and contains 3034 radio components above a signal-to-noise ratio (S/N) of 5. ELAIS-S1 covers 2.7 deg^2 and contains 2084 radio components above an S/N of 5 (Franzen et al. 2015). The images of CDFS and ELAIS-S1 have angular resolutions of 16 by 7 and 12 by 8 arcsec respectively, with pixel sizes of 1.5 arcsec pixel^{-1}. Table 1 summarizes catalogues that contain cross-identifications of radio components in ATLAS with host galaxies in SWIRE. In this work, we train methods on CDFS[3] and test these methods on both CDFS and ELAIS-S1. This ensures our methods are transferable to different areas of the sky observed by the same telescope as will be the case for EMU.

[1] https://radio.galaxyzoo.org

[2] This will be incorrect if the unresolved components are actually compact lobes or hotspots, but determining which components correspond to unique radio sources is outside the scope of this paper.

[3] RGZ only contains CDFS sources and so we cannot train methods on ELAIS-S1.



(a) Two compact components, each a compact source.

(b) One resolved component and resolved source.

(c) Three resolved components comprising one resolved source.

Figure 1. Examples showing key definitions of radio emission regions used throughout this paper. Compact and resolved components are defined by equation (1).

Table 1. Catalogues of ATLAS/SWIRE cross-identifications for the CDFS and ELAIS-S1 fields. The method used to generate each catalogue is shown, along with the number of radio components cross-identified in each field.

Catalogue	Method	CDFS	ELAIS-S1
Norris et al. (2006)	Manual	784	0
Middelberg et al. (2008)	Manual	0	1366
Fan et al. (2015)	Bayesian models	784	0
Weston et al. (2018)	Likelihood ratio	3078	2113
Wong et al. (in preparation)	Crowdsourcing	2460	0

2.2 SWIRE

SWIRE is a wide-area infrared survey at the four IRAC wavelengths 3.6, 4.5, 5.8, and 8.0 μm (Lonsdale et al. 2003; Surace et al. 2005). It covers eight fields, including CDFS and ELAIS-S1. SWIRE is the source of infrared observations for cross-identification with ATLAS. SWIRE has catalogued 221 535 infrared objects in CDFS and 186 059 infrared objects in ELAIS-S1 above an S/N of 5.

2.3 Radio Galaxy Zoo

RGZ asks volunteers to cross-identify radio components with their infrared host galaxies. There are a total of 2460 radio components in RGZ sourced from ATLAS observations of CDFS. These components are cross-identified by RGZ participants with host galaxies detected in SWIRE. A more detailed description can be found in Banfield et al. (2015) and a full description of how the RGZ catalogue used in this work[4] is generated can be found in Wong et al.(in preparation).

The ATLAS CDFS radio components that appear in RGZ are drawn from a pre-release version of the third data release of ATLAS by Franzen et al. (2015). In this release, each radio component was fit with a 2D Gaussian. Depending on the residual of the fit, more than one Gaussian may be fit to one region of radio emission. Each of these Gaussian fits is listed as a radio component in the ATLAS component catalogue. The brightest radio component from the multiple-Gaussian fit is called the 'primary component'. If there was only one Gaussian fit then this Gaussian is the primary component. Each primary component found in the ATLAS component catalogue appears in RGZ. Non-primary components may appear within the image of a primary component, but do not have their own entry in RGZ. We will henceforth only discuss the primary components.

[4]The RGZ Data Release 1 catalogue will only include cross-identifications for which over 65 per cent of volunteers agree. However, we use a preliminary catalogue containing volunteer cross-identifications for all components.

3 METHOD

The aim of this paper is to express cross-identification in a form that will allow us to apply standard machine learning tools and methods. We use an approach from computer vision to cast cross-identification as binary classification.

3.1 Cross-identification as binary classification

We propose a two-step method for host galaxy cross-identification which we will describe now. Given a radio component, we want to find the corresponding host galaxy. The input is a 2 arcmin × 2 arcmin radio image of the sky centred on a radio component and potentially other information about objects in the image (such as the redshift or infrared colour). Images at other wavelengths (notably infrared) might be useful, but we defer this for now as it complicates the task. We chose a 2 arcmin × 2 arcmin image to match the size of the images used by RGZ. To avoid solving the separate task of identifying which radio components are associated with the same source, we assume that each radio image represents a single extended source.[5] Radio cross-identification can then be formalized as follows: given a radio image centred on a radio component, locate the host galaxy of the source containing this radio component. This is a standard computer vision problem called 'object detection', and we apply a common technique called a 'sliding-window' (Rowley, Baluja & Kanade 1996).

In sliding-window object detection, we want to find an object in an image. We develop a function to score each location in the image such that the highest scored location coincides with the desired object (equation 1). Square image cutouts called 'windows' are taken

[5]Limitations of this assumption are discussed in Section 3.2.

centred on each location and these windows are used to represent that location in our scoring function. To find the infrared host galaxy, we choose the location with the highest score. To improve the efficiency of this process when applied to cross-identification, we only consider windows coincident with infrared sources detected in SWIRE. We call these infrared sources 'candidate host galaxies'. For this paper, there is no use in scoring locations without infrared sources as that would not lead to a host identification anyway. Using candidate host galaxies instead of pixels also allows us to include ancillary information about the candidate host galaxies, such as their infrared colours and redshifts. We refer to the maximum distance a candidate host galaxy can be separated from a radio component as the 'search radius' and take this radius to be 1 arcmin. To score each candidate host galaxy, we use a 'binary classifier', which we will define now.

Algorithm 1: Cross-identifying a radio component given a radio image of the component, a catalogue of infrared candidate host galaxies and a binary classifier. σ is a parameter of the method.

Data:
A 2×2 arcmin radio image of a radio component
A set of infrared candidate host galaxies $\mathcal{G}$
A binary classifier $f : \mathbb{R}^k \to \mathbb{R}$
Result: A galaxy $g \in \mathcal{G}$
$max \leftarrow -\infty$;
$host \leftarrow \emptyset$;
for $g \in \mathcal{G}$ **do**
  $\boldsymbol{x} \leftarrow$ a k-dimensional vector representation of g (Section 3.3);
  $d \leftarrow$ distance between g and the radio component;
  $score \leftarrow f(x) \times \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{d^2}{2\sigma^2}\right)$; (0.1)
  if $score > max$ **then**
    $max \leftarrow score$;
    $host \leftarrow g$;
  end
end
return $host$

Binary classification is a common method in machine learning where objects are to be assigned to one of two classes, called the 'positive' and 'negative' classes. This assignment is represented by the probability that an object is in the positive class. A 'binary classifier' is a function mapping from an object to such a probability. Our formulation of cross-identification is equivalent to binary classification of candidate host galaxies: the positive class represents host galaxies, the negative class represents non-host galaxies, and to cross-identify a radio component, we find the candidate host galaxy maximizing the positive class probability. In other words, the binary classifier is exactly the sliding-window scoring function. We therefore split cross-identification into two separate tasks: the 'candidate classification task' where, given a candidate host galaxy, we wish to determine whether it is a host galaxy of *any* radio component; and the 'cross-identification task' where, given a specific radio component, we wish to find its host galaxy. The candidate classification task is a traditional machine learning problem which results in a binary classifier. To avoid ambiguity and recognize that the values output by a binary classifier are not true probabilities, we will refer to the outputs of the binary classifier as 'scores' in line with the sliding-window approach described above. The cross-identification task maximixes over scores output by this classifier. Our approach

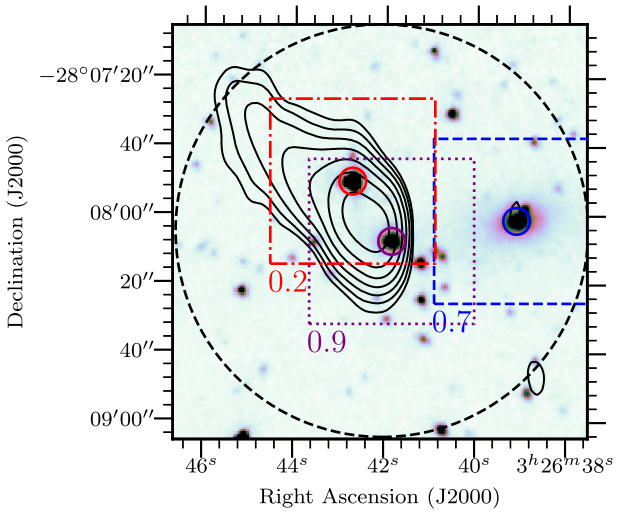


Figure 2. An example of finding the host galaxy of a radio source using our sliding-window method. The background image is a 3.6 μm image from SWIRE. The contours show ATLAS radio data and start at 4σ, increasing geometrically by a factor of 2. Boxes represent 'windows' centred on candidate host galaxies, which are circled. The pixels in each window are used to represent the candidate that the window is centred on. The scores of each candidate would be calculated by a binary classifier using the window as input, and these scores are shown below each window. The scores shown are for illustration only. In this example, the galaxy coincident with the centre window would be chosen as the host galaxy, as this window has the highest score. The dashed circle shows the 1 arcmin radius from which candidate host galaxies are selected. For clarity, not all candidate host galaxies are shown.

is illustrated in Fig. 2 and described in Algorithm 1. We refer to the binary classifier scoring a candidate host galaxy as f. To implement f as a function that accepts candidate host galaxies as input, we need to represent candidate host galaxies by vectors. We describe this in Section 3.3. There are many options for modelling f. In this paper, we apply three different models: logistic regression, random forests, and convolutional neural networks (CNNs).

We cross-identify each radio component in turn. The classifier f provides a score for each candidate host galaxy. This score indicates how much the candidate looks like a host galaxy, independent of which radio component we are currently cross-identifying. If there are other nearby host galaxies, then multiple candidate hosts may have high scores (e.g. Fig. 3). This difficulty is necessary – a classifier with dependence on radio object would be impossible to train. We need multiple positive examples (i.e. host galaxies) to train a binary classifier, but for any specific radio component there is only one host galaxy. As a result, the candidate classification task aims to answer the general question of whether a given galaxy is the host galaxy of *any* radio component, while the cross-identification task attempts to cross-identify a *specific* radio component. To distinguish between candidate host galaxies with high scores, we weight the scores by a Gaussian function of angular separation between the candidates and the radio component. The width of the Gaussian, σ, controls the influence of the Gaussian on the final cross-identification. When σ is small, our approach is equivalent to a nearest-neighbours approach where we select the nearest infrared object to the radio component as the host galaxy. In the limit where $\sigma \to \infty$, we maximize the score output by the classifier as above. We take $\sigma = 30$ arcsec as this was the best value found by a grid search. Note that the optimum width will depend on the density of

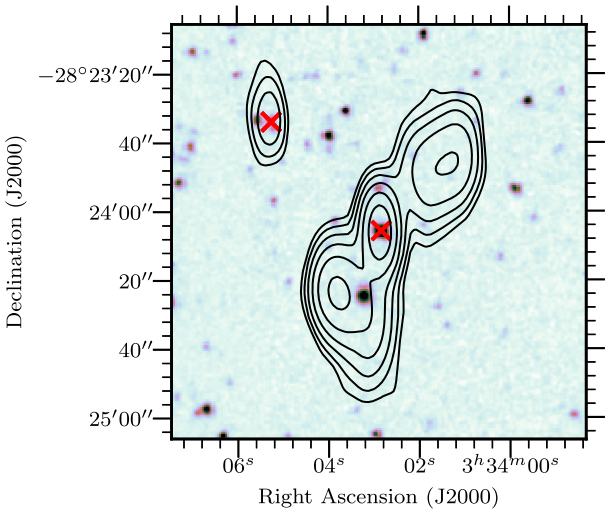


Figure 3. A 2-arcsec-wide radio image centred on ATLAS3_J033402.87-282405.8C. This radio source breaks the assumption that there are no other radio sources within 1 arcmin of the source. Another radio source is visible to the upper left. Host galaxies found by RGZ volunteers are shown by crosses. The background image is a 3.6 μm image from SWIRE. The contours show ATLAS radio data and start at 4σ, increasing geometrically by a factor of 2.

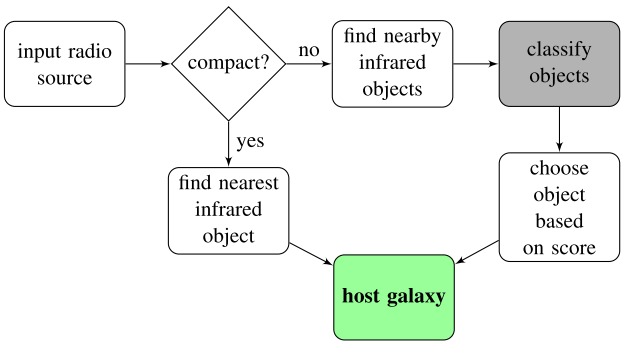


Figure 4. Our cross-identification method once a binary classifier has been trained. As input, we accept a radio component. If the component is compact, we assume it is a compact source and select the nearest infrared object as the host galaxy. If the component is resolved, we use the binary classifier to score all nearby infrared objects and select the highest scored object as the host galaxy. Compact and resolved components are defined in equation (1).

radio sources on the sky, the angular separation of the host galaxy and its radio components and the angular resolution of the survey.

We can improve upon this method by cross-identifying compact radio sources separately from extended sources, as compact sources are much easier to cross-identify. For a compact source, the nearest SWIRE object may be identified as the host galaxy (a *nearest-neighbours* approach), or a more complex method such as likelihood ratios may be applied (see Weston et al. 2018). We cross-identify compact sources separately in our pipeline and this process is shown in Fig. 4.

3.2 Limitations of our approach

We make a number of assumptions to relate the cross-identification task to the candidate classification task:

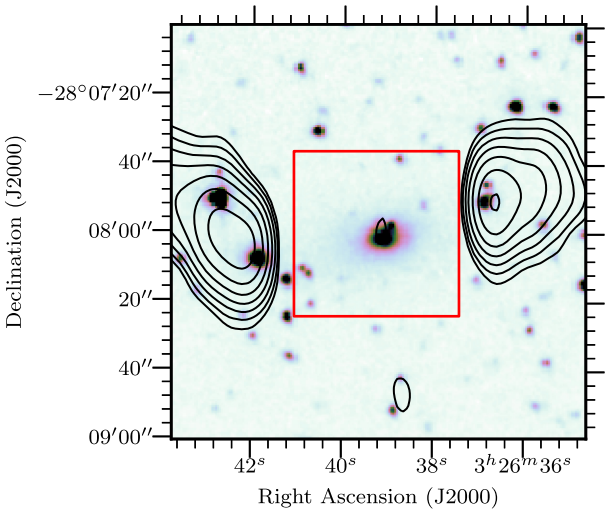


Figure 5. An example of a radio source where the window centred on the host galaxy, shown as a rectangle, does not contain enough radio information to correctly identify the galaxy as the host. The background image is a 3.6 μm image from SWIRE. The contours show ATLAS radio data and start at 4σ, increasing geometrically by a factor of 2.

(i) For any radio component, the 2 arcmin × 2 arcmin image centred on the component contains components of only one radio source.
(ii) For any radio component, the 2 arcmin × 2 arcmin image centred on the component contains all components of this source.
(iii) The host galaxy of a radio component is within the 1 arcmin search radius around the component, measured from the centre of the Gaussian fit.
(iv) The host galaxy of a radio component is closer on the sky to the radio component than the host galaxy of any other radio component.
(v) The host galaxy appears in the SWIRE catalogue.

These assumptions limit the effectiveness of our approach, regardless of how accurate our binary classifier may be. Examples of radio sources that break these respective assumptions are:

(i) A radio source less than 1 arcmin away from another radio source.
(ii) A radio source with an angular size greater than 2 arcmin.
(iii) A radio source with a component greater than 1 arcmin away from the host galaxy.
(iv) A two-component radio source with another host galaxy between a component and the true host galaxy.
(v) An infrared-faint radio source (as in Collier et al. 2014).

The main limitations are problems of scale in choosing the candidate search radius and the size of the windows representing candidates. If the search radius is too small, we may not consider the host galaxy as a candidate. If the search radius is too large, we may consider multiple host galaxies (though this is mostly mitigated by the Gaussian weighting). If the window is too small, radio emission may extend past the edges of the window and we may miss critical information required to identify the galaxy as a host galaxy. If the window is too large, then irrelevant information will be included and it may be difficult or computationally expensive to score. We chose a window size of 32 × 32 pixels, corresponding to approximately 48 arcsec × 48 arcsec in ATLAS. This is shown as squares in Figs 2 and 5. These kinds of size problems are difficult even for

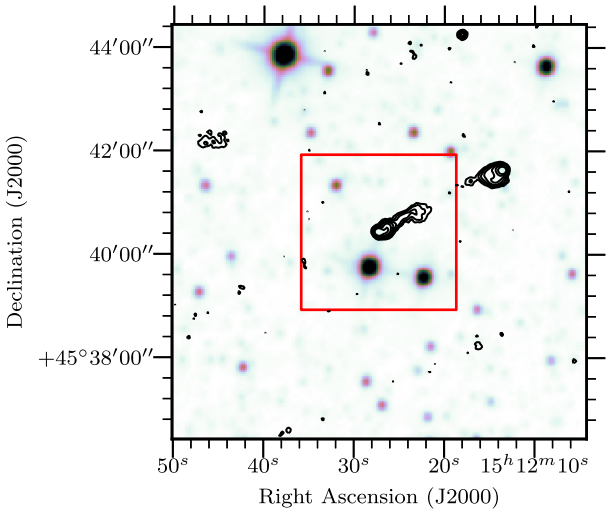


Figure 6. A 8-arcmin-wide radio image from FIRST, centred on FIRST J151227.2+454026. The 3-arcmin-wide red box indicates the boundaries of the image of this radio component shown to volunteers in RGZ. This radio source breaks our assumption that the whole radio source is visible in the chosen radius. As one of the components of the radio source is outside of the image, a volunteer (or automated algorithm) looking at the 3-arcmin-wide image may be unable to determine that this is a radio double or locate the host galaxy. The background image is a 3.4 μm image from *WISE*. The contours show FIRST radio data, starting at 4σ and increasing geometrically by a factor of 2.

non-automated methods as radio sources can be extremely wide – for example, RGZ found a radio giant that spanned over three different images presented to volunteers and the full source was only cross-identified by the efforts of citizen scientists (Banfield et al. 2015). An example of a radio image where part of the radio source is outside the search radius is shown in Fig. 6.

In weighting the scores by a Gaussian function of angular separation, we implicitly assume that the host galaxy of a radio component is closer to that radio component than any other host galaxy. If this assumption is not true then the incorrect host galaxy may be identified, though this is rare.

We only need to require that the host galaxy appears in SWIRE to incorporate galaxy-specific features (Section 3.3) and to improve efficiency. Our method is applicable even when host galaxies are not detected in the infrared by considering every pixel of the radio image as a candidate location as would be done in the original computer vision approach. If the host galaxy location does not correspond to an infrared source, the radio source would be classified as infrared-faint.

Our assumptions impose an upper bound on how well we can cross-identify radio sources. We estimate this upper bound in Section 4.1.

3.3 Feature vector representation of infrared sources

Inputs to binary classifiers must be represented by an array of real values called feature vectors. We therefore need to choose a feature vector representation of our candidate host galaxies. Candidate hosts are sourced from the SWIRE catalogue (Section 2.2). We represent each candidate host with 1034 real-valued features, combining the windows centred on each candidate (Section 3.1) with ancillary infrared data from the SWIRE catalogue. For a given candidate host, these features are:

(i) the 6 base-10 logarithms of the ratios of fluxes of the candidate host at the four IRAC wavelengths (the 'colours' of the candidate);
(ii) the flux of the host at 3.6 μm;
(iii) the stellarity index of the host at both 3.6 and 4.5 μm;
(iv) the radial distance between the candidate host and the nearest radio component in the ATLAS catalogue; and
(v) a 32 × 32 pixel image from ATLAS (approximately 48 arcsec × 48 arcsec), centred on the candidate host (the window).

The infrared colours provide insight into the properties of the candidate host galaxy (Grant 2011). The 3.6 and 4.5 μm fluxes trace both galaxies with faint polycyclic aromatic hydrocarbon (PAH) emission (i.e. late-type, usually star-forming galaxies) and elliptical galaxies dominated by old stellar populations. The 5.8 μm flux selects galaxies where the infrared emission is dominated by non-equilibrium emission of dust grains due to active galactic nuclei, while the 8.0 μm flux traces strong PAH emission at low redshift (Sajina, Lacy & Scott 2005). The stellarity index is a value in the SWIRE catalogue that represents how likely the object is to be a star rather than a galaxy (Surace et al. 2005). It was estimated by a neural network in SEXTRACTOR (Bertin & Arnouts 1996).

We use the 32 × 32 pixels of each radio window as independent features for all binary classification models, with the CNN automatically extracting features that are relevant. Other features of the radio components may be used instead of just relying on the pixel values, but there has been limited research on extracting such features: Proctor (2006) describes hand-selected features for radio doubles in FIRST, and Aniyan & Thorat (2017) and Lukic et al. (2018) make use of deep CNNs which automatically extract features as part of classification. A more comprehensive investigation of features is a good avenue for potential improvement in our pipeline but this is beyond the scope of this initial study.

3.4 Binary classifiers

We use three different binary classification models: logistic regression, CNNs, and random forests. These models cover three different approaches to machine learning. Logistic regression is a probabilistic binary classification model. It is linear in the feature space and outputs the probability that the input has a positive label (Bishop 2006, chap. 4). CNNs are biologically inspired prediction models with image inputs. They have recently produced good results on large image-based data sets in astronomy (e.g. Dieleman, Willett & Dambre 2015;Lukic et al. 2018). Random forests are an ensemble of decision trees (Breiman 2001). They consider multiple subsamples of the training set, where each bootstrap subsample is sampled with replacement from the training set. To classify a new data point, the random forest takes the weighted average of all classifications produced by each decision tree.

Further details and background of these models are presented in Appendix A.

3.5 Labels

The RGZ and Norris et al. (2006) cross-identification catalogues must be converted to binary labels for infrared objects so that they can be used to train binary classifiers. There are two challenges with this conversion:

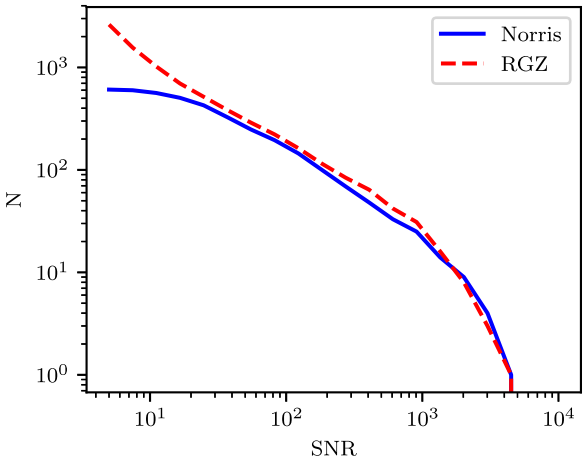


Figure 7. Cumulative number of radio components (N) in the expert (Norris) and RGZ training sets with different S/Ns.

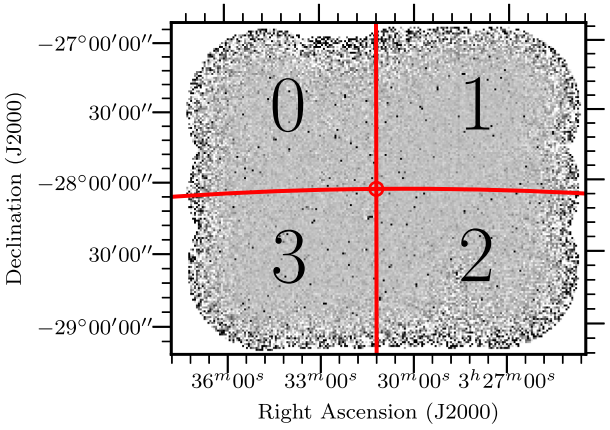


Figure 8. CDFS field training and testing quadrants labelled 0 –3. The central dot is located at $\alpha = 03^{\rm h}31^{\rm m}12^{\rm s}$ and $\delta = -28^\circ06'00''$. The quadrants were chosen such that there are similar numbers of radio sources in each quadrant.

(i) We can only say that an object is *a* host galaxy, not which radio object it is associated with, and
(ii) We cannot disambiguate between non-host infrared objects and host galaxies that were not in the cross-identification catalogue.

We use the Gaussian weighting described in Section 3.1 to address the first issue. The second issue is known as a 'positive-unlabelled' classification problem, which is a binary classification problem where we only observe labels for the positive class. We treat unlabelled objects as negative examples following Menon et al. (2015). That is, we make the naïve assumption that any infrared object in the SWIRE catalogue not identified as a host galaxy in a cross-identification catalogue is not a host galaxy at all.

We first generate positive labels from a cross-identification catalogue. We decide that if an infrared object is listed in the catalogue, then it is assigned a positive label as a host galaxy. We then assign every other galaxy a negative label. This has some problems – an example is that if the cross-identification catalogue did not include a radio object (e.g. it was below the S/N) then the host galaxy of that radio object would receive a negative label. This occurs with Norris et al. (2006) cross-identifications, as these are associated with the first data release of ATLAS. The first data release went to a 5σ flux density level of $S_{1.4} \geq 200\,\mu$Jy beam^{-1} (Norris et al. 2006), compared to $S_{1.4} \geq 85\,\mu$Jy beam^{-1} for the third data release used by RGZ (Franzen et al. 2015). The labels from Norris et al. (2006) may therefore disagree with labels from RGZ even if they are both plausible. The difference in training set size at different flux cut-offs is shown in Fig. 7. We train and test our binary classifiers on infrared objects within a 1 arcmin radius of an ATLAS radio component.

3.6 Experimental setup

We trained binary classifiers on infrared objects in the CDFS field using two sets of labels. One label set was derived from RGZ cross-identifications and the other was derived from the Norris et al. (2006) cross-identification catalogue. We refer to these as the 'RGZ labels' and the 'expert labels' respectively. We divided the CDFS field into four quadrants for training and testing. The quadrants were divided with a common corner at $\alpha = 03^{\rm h}31^{\rm m}12^{\rm s}$ and $\delta = -28^\circ06'00''$ as shown in Fig. 8. For each trial, one quadrant was used to extract test examples and the other three quadrants were used for training examples.

We further divided the radio components into compact and resolved. Compact components are cross-identified by fitting a 2D Gaussian (as in Norris et al. 2006) and we would expect any machine learning approach for host cross-identification to attain high accuracy on this set. A radio component was considered resolved if

$$\ln\left(\frac{S_{\rm int}}{S_{\rm peak}}\right) > 2\sqrt{\left(\frac{\sigma_{S_{\rm int}}}{S_{\rm int}}\right)^2 + \left(\frac{\sigma_{S_{\rm peak}}}{S_{\rm peak}}\right)^2}, \quad (1)$$

where $S_{\rm int}$ is the integrated flux density, $S_{\rm peak}$ is the peak flux density, $\sigma_{S_{\rm int}}$ is the uncertainty in integrated flux density, and $\sigma_{S_{\rm peak}}$ is the uncertainty in peak flux density (following Franzen et al. 2015).

Candidate hosts were selected from the SWIRE catalogue. For a given subset of radio components, all SWIRE objects within 1 arcmin of all radio components in the subset were added to the associated SWIRE subset. In results for the candidate classification task, we refer to SWIRE objects within 1 arcmin of a compact radio component as part of the 'compact set', and SWIRE objects within 1 arcmin of a resolved radio component as part of the 'resolved set'.

To reduce bias in the testing data due to the expert labels being generated from a shallower data release of ATLAS, a SWIRE object was only included in the test set if it was within 1 arcmin of a radio object with a SWIRE cross-identification in both the Norris et al. (2006) catalogue and the RGZ catalogue.

Each binary classifier was trained on the training examples and used to score the test examples. These scores were thresholded to generate labels which could be directly compared to the expert labels. We then computed the 'balanced accuracy' of these predicted labels. Balanced accuracy is the average of the accuracy on the positive class and the accuracy on the negative class, and is not sensitive to class imbalance. The candidate classification task has highly imbalanced classes – in our total set of SWIRE objects within 1 arcmin of an ATLAS object, only 4 per cent have positive labels. Our threshold was chosen to maximize the balanced accuracy on predicted labels of the training set. Only examples within 1 arcmin of ATLAS objects in the first ATLAS data release (Norris et al. 2006) were used to compute balanced accuracy, as these were the only ATLAS objects with expert labels.

We then used the scores to predict the host galaxy for each radio component cross-identified by both Norris et al. (2006) and RGZ. We followed Algorithm 1: the score of each SWIRE object within 1 arcmin of a given radio component was weighted by a Gaussian

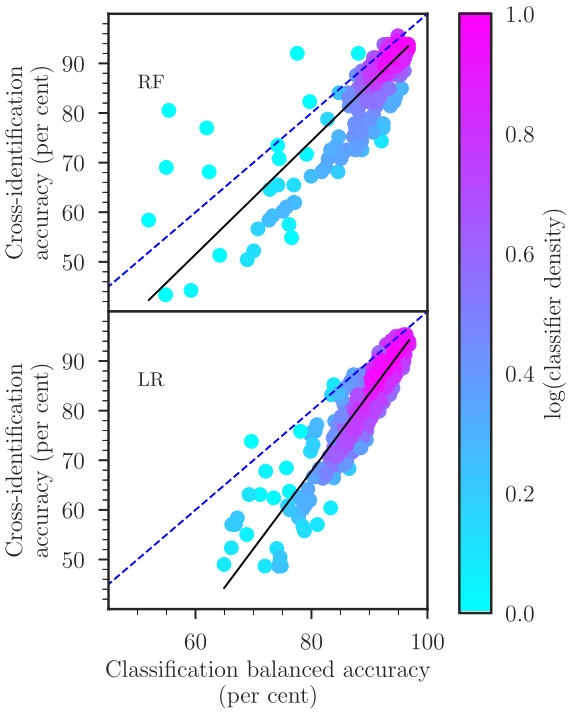


Figure 9. Balanced accuracy on the candidate classification task plotted against accuracy on the cross-identification task. 'RF' indicates results from random forests, and 'LR' indicates results from logistic regression. Binary classifiers were trained on random, small subsets of the training data to artificially restrict their accuracies. Colour shows the density of points on the plot estimated by a Gaussian kernel density estimate. The solid lines indicate the best linear fit; these fits have $R^2 = 0.92$ for logistic regression and $R^2 = 0.87$ for random forests. The dashed line shows the line where cross-identification accuracy and candidate classification accuracy are equal. We did not include CNNs in this test, as training them is very computationally expensive. There are 640 trials shown per classification model. These results exclude binary classifiers with balanced accuracies less than 51 per cent, as these are essentially random.

function of angular separation from the radio component and the object with the highest weighted score was chosen as the host galaxy. The cross-identification accuracy was then estimated as the fraction of the predicted host galaxies that matched the Norris et al. (2006) cross-identifications.

4 RESULTS

In this section, we present accuracies of our method trained on CDFS and applied to CDFS and ELAIS-S1, as well as results motivating our accuracy measures and estimates of upper and lower bounds for cross-identification accuracy using our method.

4.1 Application to ATLAS–CDFS

We can assess trained binary classifiers either by their performance on the candidate classification task or by their performance on the cross-identification task when used in our method. Both performances are useful: performance on the candidate classification task provides a robust and simple way to compare binary classifiers without the limitations of our specific formulation, and performance on the cross-identification task can be compared with other cross-identification methods. We therefore report two sets of accuracies: balanced accuracy for the galaxy classification task and accuracy

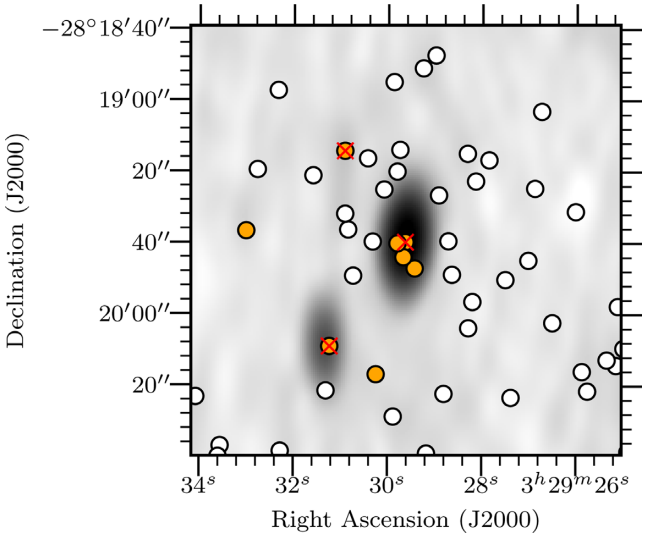


Figure 10. Predicted host galaxies in the candidate classification task for ATLAS3 J032929.61−281938.9. The background image is an ATLAS radio image. RGZ host galaxies are marked by crosses. SWIRE candidate host galaxies are circles coloured by the score output by a logistic regression binary classifier. The scores are thresholded to obtain labels, as when we compute balanced accuracy. Orange circles have been assigned a 'positive' label by a logistic regression binary classifier and white otherwise. Note that there are more predicted host galaxies than there are radio components, so not all of the predicted host galaxies would be assigned as host galaxies in the cross-identification task.

for the cross-identification task. These accuracy measures are correlated and we show this correlation in Fig. 9. Fitting a line of best fit with SCIPY gives $R^2 = 0.92$ for logistic regression and $R^2 = 0.87$ for random forests. While performance on the candidate classification task is correlated with performance on the cross-identification task, balanced accuracy does not completely capture the effectiveness of a binary classifier applied to the cross-identification task. This is because while our binary classifiers output real-valued scores, these scores are thresholded to compute the balanced accuracy. In the candidate classification task, the binary classifier only needs to ensure that host galaxies are scored higher than non-host galaxies. This means that after thresholding there can be many 'false positives' that do not affect cross-identification. An example of this is shown in Fig. 10, where the classifier has identified eight 'host galaxies'. However, there are only three true host galaxies in this image – one per radio component – and so in the cross-identification task, only three of these galaxies will be identified as hosts.

In Fig. 11, we plot the balanced accuracies of our classification models on the candidate classification task and the cross-identification accuracies of our method using each of these models. Results are shown for both the resolved and compact sets. For comparison, we also plot the cross-identification accuracy of RGZ and a nearest-neighbours approach, as well as estimates for upper and lower limits on the cross-identification accuracy. We estimate the upper limit on performance by assigning all true host galaxies a score of 1 and assigning all other candidate host galaxies a score of 0. This is equivalent to 'perfectly' solving the candidate classification task and so represents the best possible cross-identification performance achievable with our method. We estimate the lower limit on performance by assigning random scores to each candidate host galaxy. We expect any useful binary classifier to produce better results than this, so this represents the lowest expected

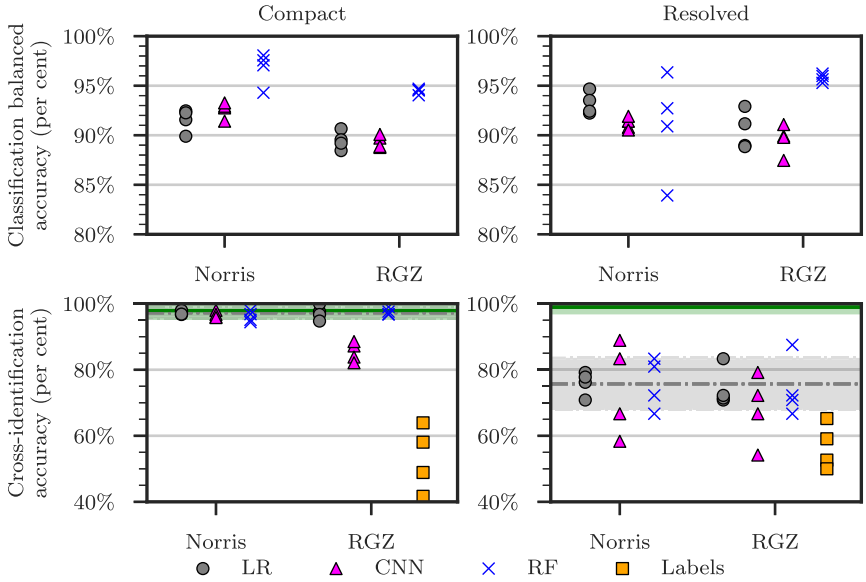


Figure 11. Performance of our method with logistic regression ('LR'), convolutional neural networks ('CNN'), and random forest ('RF') binary classifiers. 'Norris' indicates the performance of binary classifiers trained on the expert labels and 'RGZ' indicates the performance of binary classifiers trained on the Radio Galaxy Zoo labels. One point is shown per binary classifier per testing quadrant. The training and testing sets have been split into compact (left) and resolved (right) objects. Shown for comparison is the accuracy of the RGZ consensus cross-identifications on the cross-identification task, shown as 'Labels'. The cross-identification accuracy attained by a perfect binary classifier is shown by a solid green line, and the cross-identification accuracy of nearest-neighbours approach is shown by a dashed grey line. The standard deviation of these accuracies across the four CDFS quadrants is shown by the shaded area. Note that the pipeline shown in Fig. 4 is not used for these results.

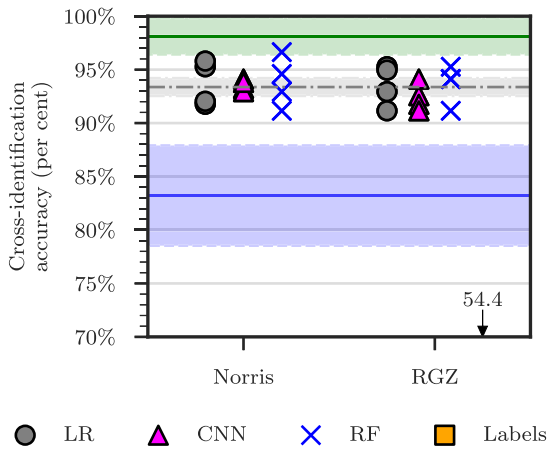


Figure 12. Performance of our approach using different binary classifiers on the cross-identification task. Markers and lines are as in Fig. 11. The blue solid line indicates the performance of a random binary classifier and represents the minimum accuracy we expect to obtain. The standard deviation of this accuracy across 25 trials and four quadrants is shaded. The accuracy of RGZ on the cross-identification task is below the axis and is instead marked by an arrow with the mean accuracy. Note that the pipeline shown in Fig. 4 is used here, so compact objects are cross-identified in the same way regardless of binary classifier model.

cross-identification performance. The upper estimates, lower estimates, and nearest-neighbour accuracy are shown as horizontal lines in Fig. 11.

In Fig. 12, we plot the performance of our method using different binary classification models, as well as the performance of RGZ, nearest-neighbours, and the perfect and random binary classifiers, on the full set of ATLAS DR1 radio components using the pipeline in Fig. 4. The accuracy associated with each classification model and training label set averaged across all four quadrants is shown in Appendix B.

Table 2. Number of compact and resolved radio objects in each CDFS quadrant. RGZ has more cross-identifications than the expert catalogue (Norris et al. 2006) provides as it uses a deeper data release of ATLAS, and so has more objects in each quadrant for training.

Quadrant	Compact	Resolved	Compact (RGZ)	Resolved (RGZ)
0	126	24	410	43
1	99	21	659	54
2	61	24	555	57
3	95	18	631	51
Total	381	87	2255	205

Differences between accuracies across training labels are well within one standard deviation computed across the four quadrants, with CNNs on compact objects as the only exception. The spread of accuracies is similar for both sets of training labels, with the exception of random forests. The balanced accuracies of random forests trained on expert labels have a considerably higher spread than those trained on RGZ labels, likely because of the small size of the expert training set – there are less than half the number of objects in the expert-labelled training set than the number of objects in the RGZ-labelled training set (Table 2).

RGZ-trained methods significantly outperform RGZ cross-identifications. Additionally, despite poor performance of RGZ on the cross-identification task, methods trained on these cross-identifications still perform comparably to those trained on expert labels. This is because incorrect RGZ cross-identifications can be thought of as a source of noise in the labels which is 'averaged out' in training. This shows the usefulness of crowd-sourced training data, even when the data are noisy.

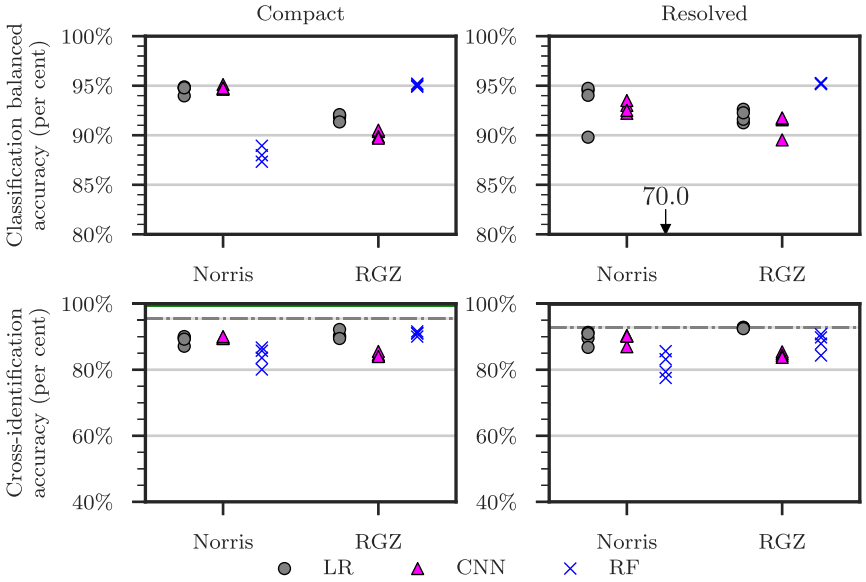


Figure 13. Performance of different classification models trained on CDFS and tested on resolved and compact sources in ELAIS-S1. Points represent classification models trained on different quadrants of CDFS, with markers, lines, and axes as in Fig. 11. The balanced accuracy of expert-trained random forest binary classifiers falls below the axis and the corresponding mean accuracy is shown by an arrow. The estimated best attainable accuracy is almost 100 per cent.

Our method performs comparably to a nearest-neighbours approach. For compact objects, this is to be expected – indeed, nearest-neighbours attains nearly 100 per cent accuracy on the compact test set. Our results do not improve on nearest-neighbours for resolved objects. However, our method does allow for improvement on nearest-neighbours with a sufficiently good binary classifier: a 'perfect' binary classifier attains nearly 100 per cent accuracy on resolved sources. This shows that our method may be useful provided that a good binary classifier can be trained. The most obvious place for improvement is in feature selection: we use pixels of radio images directly and these are likely not conducive to good performance on the candidate classification task. CNNs, which are able to extract features from images, *should* work better, but these require far more training data than the other methods we have applied and the small size of ATLAS thus limits their performance.

We noted in Section 3.5 that the test set of expert labels, derived from the initial ATLAS data release, was less deep than the third data release used by RGZ and this paper, introducing a source of label noise in the testing labels. Specifically, true host galaxies may be misidentified as non-host galaxies if the associated radio source was below the 5 S/N limit in ATLAS DR1 but not in ATLAS DR3. This has the effect of reducing the accuracy for RGZ-trained classifiers.

We report the scores predicted by each classifier for each SWIRE object in Appendix C and the predicted cross-identification for each ATLAS object in Appendix D. Scores reported for a given object were predicted by binary classifiers tested on the quadrant containing that object. The reported scores are not weighted.

In Fig. E1, we show five resolved sources where the most classifiers disagreed on the correct cross-identification.

4.2 Application to ATLAS–ELAIS-S1

We applied the method trained on CDFS to perform cross-identification on the ELAIS-S1 field. Both CDFS and ELAIS-S1 were imaged by the same radio telescope to similar sensitivities and angular resolution for the ATLAS survey. We can use the SWIRE cross-identifications made by Middelberg et al. (2008) to derive another set of expert labels, and hence determine how accurate our method is. If our method generalizes well across different parts of the sky, then we expect CDFS-trained classifiers to have comparable performance between ELAIS-S1 and CDFS. In Fig. 13, we plot the performance of CDFS-trained classification models on the candidate classification task and the performance of our method on the cross-identification task using these models. We also plot the cross-identification accuracy of a nearest-neighbours approach.[6] In Fig. 14, we plot the performance of our method on the full set of ELAIS-S1 ATLAS DR1 radio components using the pipeline in Fig. 4. We list the corresponding accuracies in Appendix B.

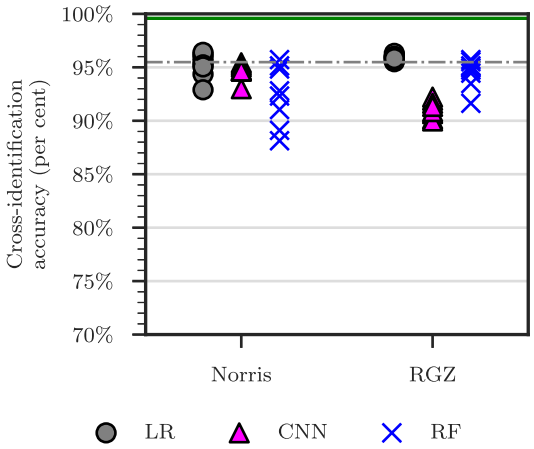


Figure 14. Performance of different classifiers trained on CDFS and tested on ELAIS-S1. Markers are as in Fig. 12 and horizontal lines are as in Fig. 13. Note that the pipeline shown in Fig. 4 is used here, so compact objects are cross-identified in the same way regardless of binary classifier model.

[6]We cannot directly compare our method applied to ELAIS-S1 with RGZ, as RGZ does not include ELAIS-S1.

Cross-identification results from ELAIS-S1 are similar to those for CDFS, showing that our method trained on CDFS performs comparably well on ELAIS-S1. However, nearest-neighbours outperforms most methods on ELAIS-S1. This is likely because there is a much higher percentage of compact objects in ELAIS-S1 than in CDFS. The maximum achievable accuracy we have estimated for ELAIS-S1 is very close to 100 per cent, so (as for CDFS) a very accurate binary classifier would outperform nearest-neighbours.

One interesting difference between the ATLAS fields is that random forests trained on expert labels perform well on CDFS but poorly on ELAIS-S1. This is not the case for logistic regression or CNNs trained on expert labels, nor is it the case for random forests trained on RGZ. We hypothesize that this is because the ELAIS-S1 cross-identification catalogue (Middelberg et al. 2008) labelled fainter radio components than the CDFS cross-identification catalogue (Norris et al. 2006) due to noise from the very bright source ATCDFS_J032836.53-284156.0 in CDFS. Classifiers trained on CDFS expert labels may thus be biased toward brighter radio components compared to ELAIS-S1. RGZ uses a preliminary version of the third data release of ATLAS (Franzen et al. 2015) and so classifiers trained on the RGZ labels may be less biased toward brighter sources compared to those trained on the expert labels. To test this hypothesis, we tested each classification model against test sets with an S/N cut-off. A SWIRE object was only included in the test set for a given cut-off if it was located within 1 arcmin of a radio component with an S/N above the cut-off. The balanced accuracies for each classifier at each cut-off are shown in Figs 15(a) and (b) and the distribution of test set size for each cut-off is shown in Fig. 15(c). Fig. 15(c) shows that ELAIS-S1 indeed has more faint objects in its test set than the CDFS test set, with the S/N for which the two fields reach the same test set size (approximately 34) indicated by the dashed vertical line on each plot. For CDFS, all classifiers perform reasonably well across cut-offs, with performance dropping as the size of the test set becomes small. For ELAIS-S1, logistic regression and CNNs perform comparably across all S/N cut-offs, but random forests do not. While random forests trained on RGZ labels perform comparably to other classifiers across all S/N cut-offs, random forests trained on expert labels show a considerable drop in performance below the dashed line.

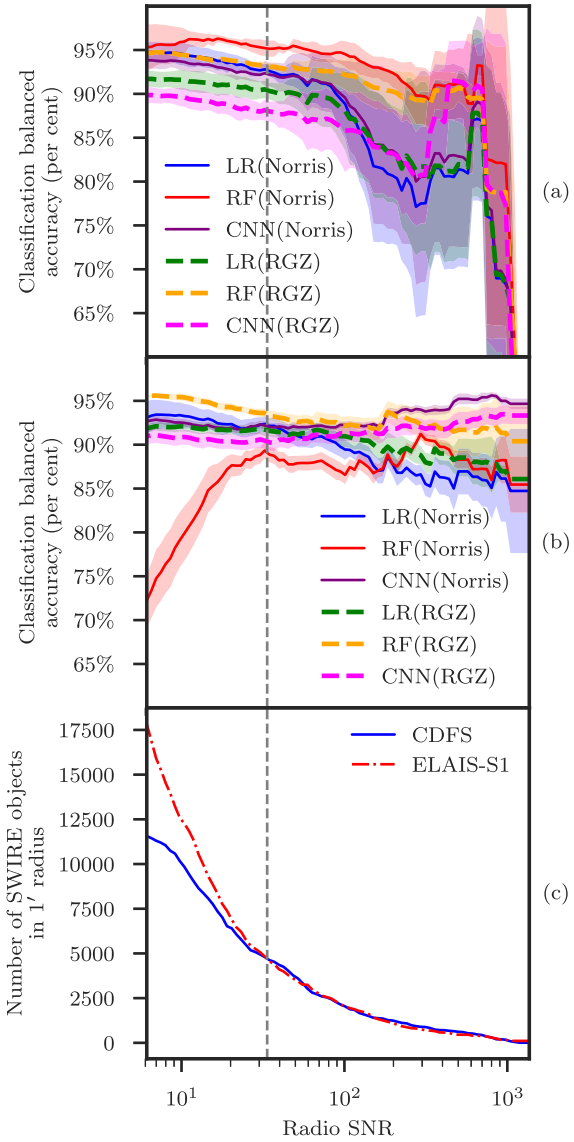


Figure 15. (a) Balanced accuracies of classifiers trained and tested on CDFS with different S/N cut-offs for the test set. A SWIRE object is included in the test set if it is within 1 arcmin of a radio component with greater S/N than the cut-off. Lines of different colour indicate different classifier/training labels combinations, where LR is logistic regression, RF is random forests, CNN is convolutional neural networks, and Norris and RGZ are the expert and Radio Galaxy Zoo label sets, respectively. Filled areas represent standard deviations across CDFS quadrants. (b) Balanced accuracies of classifiers trained on CDFS and tested on ELAIS-S1. (c) A cumulative distribution plot of SWIRE objects associated with a radio object with greater S/N than the cut-off. The grey dashed line shows the S/N level at which the number of SWIRE objects above the cut-off is equal for CDFS and ELAIS-S1. This cut-off level is approximately at an S/N of 34.

5 DISCUSSION

Based on the ATLAS sample, our main result is that it is possible to cast radio host galaxy cross-identification as a machine learning task for which standard methods can be applied. These methods can then be trained with a variety of label sets derived from cross-identification catalogues. While our methods have not outperformed nearest-neighbours, we have demonstrated that for a very accurate binary classifier, good cross-identification results can be obtained using our method. Future work could combine multiple catalogues or physical priors to boost performance.

Nearest-neighbours approaches outperform most methods we investigated, notably including RGZ. This is due to the large number of compact or partially resolved objects in ATLAS. This result shows that for compact and partially resolved objects methods that do not use machine learning such as a nearest-neighbours approach or likelihood ratio (Weston et al. 2018) should be preferred to machine learning methods. It also shows that ATLAS is not an ideal data set for developing machine learning methods like ours. Our use of ATLAS is motivated by its status as a pilot survey for EMU, so methods developed for ATLAS should also work for EMU. New methods developed should work well with extended radio sources, but this goal is almost unsupported by ATLAS as it has very few examples of such sources. This makes both training and testing difficult – there are too few extended sources to train on and performance on such a small test set may be unreliable. Larger data sets with many extended sources like FIRST exist, but these are considerably less deep than and at a different resolution to EMU, so there is no reason to expect methods trained on such data sets to be applicable to EMU.

The accuracies of our trained cross-identification methods generally fall far below the estimated best possible accuracy attainable using our approach, indicated by the green-shaded areas in Figs 12 and 14. The balanced accuracies attained by our binary classifiers indicate that there is significant room for improvement in classification. The classification accuracy could be improved by better model selection and more training data, particularly for CNNs. There is a huge variety of ways to build a CNN, and we have only investigated one architecture. For an exploration of different CNN architectures applied to radio astronomy, see Lukic et al. (2018). CNNs generally require more training data than other machine learning models and we have only trained our networks on a few hundred sources. We would expect performance on the classification task to greatly increase with larger training sets.

Another problem is that of the window size used to select radio features. Increasing window size would increase computational expense, but provide more information to the models. Results are also highly sensitive to how large the window size is compared to the size of the radio source we are trying to cross-identify, with large angular sizes requiring large window sizes to ensure that the features contain all the information needed to localize the host galaxy. An ideal implementation of our method would most likely represent a galaxy using radio images taken at multiple window sizes, but this is considerably more expensive.

Larger training sets, better model selection, and larger window sizes would improve performance, but only so far: we would still be bounded above by the estimated 'perfect' classifier accuracy. From this point, the performance can only be improved by addressing our broken assumptions. We detailed these assumptions in Section 3.2, and we will discuss here how our method could be adapted to avoid these assumptions. Our assumption that the host galaxy is contained within the search radius could be improved by dynamically choosing the search radius, perhaps based on the angular extent of the radio emission, or the redshift of candidate hosts. Radio morphology information may allow us to select relevant radio data and hence relax the assumption that a 1-arcmin-wide radio image represents just one, whole radio source. Finally, our assumption that the host galaxy is detected in infrared is technically not needed, as the sliding-window approach we have employed will still work even if there are no detected host galaxies – instead of classifying candidate hosts, simply classify each pixel in the radio image. The downside of removing candidate hosts is that we are no longer able to reliably incorporate host galaxy information such as colour and redshift, though this could be resolved by treating pixels as potentially undetected candidate hosts with noisy features.

We observe that RGZ-trained methods perform comparably to methods trained on expert labels. This shows that the crowdsourced labels from RGZ will provide a valuable source of training data for future machine learning methods in radio astronomy.

Compared to nearest-neighbours, cross-identification accuracy on ELAIS-S1 is lower than on CDFS. Particularly notable is that our performance on compact objects is very low for ELAIS-S1, while it was near-optimal for CDFS. These differences may be for a number of reasons. ELAIS-S1 has beam size and noise profile different from CDFS (even though both were imaged with the same telescope), so it is possible that our methods over-adapted to the beam and noise of CDFS. Additionally, CDFS contains a very bright source which may have caused artefacts throughout the field that are not present in ELAIS-S1. Further work is required to understand the differences between the fields and their effect on performance.

Fig. 15 reveals interesting behaviour of different classifier models at different flux cut-offs. Logistic regression and CNNs seem relatively independent of flux, with these models performing well on the fainter ELAIS-S1 components even when they were trained on the generally brighter components in CDFS. Conversely, random forests were sensitive to the changes in flux distribution between data sets. This shows that not all models behave similarly on radio data, and it is therefore important to investigate multiple models when developing machine learning methods for radio astronomy.

Appendix E (see Fig. E1) shows examples of incorrectly cross-identified components in CDFS. On no such component do all classifiers agree. This raises the possibility of using the level of disagreement of an ensemble of binary classifiers as a measure of the difficulty of cross-identifying a radio component, analogous to the consensus level for RGZ volunteers.

Our methods can be easily incorporated into other cross-identification methods or used as an extra data source for source detection. For example, the scores output by our binary classifiers could be used to disambiguate between candidate host galaxies selected by model-based algorithms, or used to weight candidate host galaxies while a source detector attempts to associate radio components. Our method can also be extended using other data sources: for example, information from source identification algorithms could be incorporated into the feature set of candidate host galaxies.

6 SUMMARY

We presented a machine learning approach for cross-identification of radio components with their corresponding infrared host galaxy. Using the CDFS field of ATLAS as a training set we trained our methods on expert and crowdsourced cross-identification catalogues. Applying these methods on both fields of ATLAS, we found that:

(i) Our method trained on ATLAS observations of CDFS generalized to ATLAS observations of ELAIS-S1, demonstrating that training on a single patch of sky is a feasible option for training machine learning methods for wide-area radio surveys;

(ii) Performance was comparable to nearest-neighbours even on resolved sources, showing that nearest-neighbours is useful for data sets consisting mostly of unresolved sources such as ATLAS and EMU;

(iii) RGZ-trained models performed comparably to expert-trained models and outperformed RGZ, showing that crowdsourced labels are useful for training machine learning methods for cross-identification even when these labels are noisy;

(iv) ATLAS does not contain sufficient data to train or test machine learning cross-identification methods for extended radio sources. This suggests that if machine learning methods are to be used on EMU, a larger area of sky will be required for training and testing these methods. However, existing surveys like FIRST are likely too different from EMU to expect good generalization.

While our cross-identification performance is not as high as desired, we make no assumptions on the binary classification model used in our methods and so we expect the performance to be improved by further experimentation and model selection. Our method provides a useful framework for generalizing cross-identification catalogues to other areas of the sky from the same radio survey and can be incorporated into existing methods. We have shown that citizen science can provide a useful data set for training machine learning methods in the radio domain.

ACKNOWLEDGEMENTS

This publication has been made possible by the participation of more than 11 000 volunteers in the Radio Galaxy Zoo project. Their contributions are individually acknowledged at http://rgzauthors.galaxyzoo.org. Parts of this research were conducted by the Australian Research Council Centre of Excellence for All-sky Astrophysics, through project number CE110001020. Partial support for LR was provided by U.S. National Science Foundation grants AST1211595 and 1714205 to the University of Minnesota. HA benefitted from grant 980/2016-2017 of Universidad de Guanajuato. We thank A. Tran and the reviewer for their comments on this manuscript. Radio Galaxy Zoo makes use of data products from the *Wide-field Infrared Survey Explorer* and the Very Large Array. The *Wide-field Infrared Survey Explorer* is a joint project of the University of California, Los Angeles, and the Jet Propulsion Laboratory/California Institute of Technology, funded by the National Aeronautics and Space Administration. The National Radio Astronomy Observatory is a facility of the National Science Foundation operated under cooperative agreement by Associated Universities, Inc. The figures in this work made use of ASTROPY, a community-developed core PYTHON package for Astronomy (Astropy Collaboration et al. 2013). The Australia Telescope Compact Array is part of the Australia Telescope, which is funded by the Commonwealth of Australia for operation as a National Facility managed by the CSIRO.

REFERENCES

Aniyan A. K., Thorat K., 2017, ApJS, 230, 20
Astropy Collaboration et al., 2013, A&A, 558, A33
Banfield J. K. et al., 2015, MNRAS, 453, 2326
Bertin E., Arnouts S., 1996, A&AS, 117, 393
Bishop C. M., 2006, Pattern Recognition and Machine Learning. Springer, Berlin
Breiman L., 2001, Mach. Learn., 45, 5
Chollet F. et al., 2015, Keras. Available at: https://keras.io/, last accessed on 22 June 2018
Collier J. D. et al., 2014, MNRAS, 439, 545
Cutri R. et al., 2013, Explanatory Supplement to the AllWISE Data Release Products. Available at: wise2.ipac.caltech.edu/docs/release/allwise/expsup, last accessed on 26 June 2018
Dieleman S., Willett K. W., Dambre J., 2015, MNRAS, 450, 1441
Fan D., Budavári T., Norris R. P., Hopkins A. M., 2015, MNRAS, 451, 1299
Franzen T. M. O. et al., 2015, MNRAS, 453, 4020
Gendre M. A., Wall J. V., 2008, MNRAS, 390, 819
Grant J. K., 2011, PhD thesis, University of Calgary
Grant J. K., Taylor A. R., Stil J. M., Landecker T. L., Kothes R., Ransom R. R., Scott D., 2010, ApJ, 714, 1689
Johnston S. et al., 2007, PASA, 24, 174
LeCun Y., Bottou L., Bengio Y., Haffner P., 1998, Proc. IEEE, 86, 2278
Lintott C. J. et al., 2008, MNRAS, 389, 1179
Lonsdale C. J. et al., 2003, PASP, 115, 897
Lukic V., Brüggen M., Banfield J. K., Wong O. I., Rudnick L., Norris R. P., Simmons B., 2018, MNRAS, 476, 246
Marshall P. J., Lintott C. J., Fletcher L. N., 2015, ARA&A, 53, 247
Menon A. K., Van Rooyen B., Ong C. S., Williamson R. C., 2015, in Proceedings of the 32nd International Conference on International Conference on Machine Learning - Vol. 37 (ICML'15), p. 125. Available at: http://dl.acm.org/citation.cfm?id=3045118.3045133, last accessed on 22 June 2018
Middelberg E. et al., 2008, AJ, 135, a1276
Norris R. P., 2017a, Nat. Astron., 1, 671
Norris R. P., 2017b, PASA, 34, e007
Norris R. P. et al., 2006, AJ, 132, 2409
Norris R. P. et al., 2011, PASA, 28, 215
Pedregosa F. et al., 2011, J. Mach. Learn. Res., 12, 2825
Proctor D. D., 2006, ApJS, 165, 95
Richter G. A., 1975, Astronomische Nachrichten, 296, 65
Rowley H. A., Baluja S., Kanade T., 1996, in Mozer M. C., Jordan M. I., Petsche T., eds, Advances in Neural Information Processing Systems, NIPS, p. 875
Sajina A., Lacy M., Scott D., 2005, ApJ, 621, 256
Surace J. A., Shupe D. L. , Fang F., Evans T., Alexov A., Frayer D., Lonsdale C. J., SWIRE Team, 2005, AAS, Vol. 37, AAS Meeting #207, p. 1246
Taylor A. R. et al., 2007, ApJ, 666, 201
Verheijen M. A. W., Oosterloo T. A., van Cappellen W. A., Bakker L., Ivashina M. V., van der Hulst J. M., 2008, in AIP Conf. Ser. 1035, The Evolution of Galaxies Through the Neutral Hydrogen Window, Am. Inst. Phys., New York. pp. 265
Weston S. D., Seymour N., Gulyaev S., Norris R. P., Banfield J., Vaccari M., Hopkins A. M., Franzen T. M. O., 2018, MNRAS, 473, 4523
White R. L., Becker R. H., Helfand D. J., Gregg M. D., 1997, ApJ, 475, 479

SUPPORTING INFORMATION

Supplementary data are available at *MNRAS* online.

rgz-cdfs-ms-supp.zip

Please note: Oxford University Press is not responsible for the content or functionality of any supporting materials supplied by the authors. Any queries (other than missing material) should be directed to the corresponding author for the article.

APPENDIX A: CLASSIFICATION MODELS

We use three different models for binary classification: logistic regression, CNNs, and random forests.

A1 Logistic regression

Logistic regression is linear in the feature space and outputs the probability that the input has a positive label. The model is (Bishop 2006):

$$f(\boldsymbol{x}) = \sigma(\boldsymbol{w}^T \boldsymbol{x} + b), \tag{A1}$$

where $\boldsymbol{w} \in \mathbb{R}^D$ is a vector of parameters, $b \in \mathbb{R}$ is a bias term, $\boldsymbol{x} \in \mathbb{R}^D$ is the feature vector representation of a candidate host, and $\sigma : \mathbb{R} \to \mathbb{R}$ is the logistic sigmoid function:

$$\sigma(a) = (1 + \exp(-a))^{-1}. \tag{A2}$$

The logistic regression model is fully differentiable, and the parameters $\boldsymbol{w}$ can therefore be learned using gradient-based optimization methods. We used the SCIKIT-LEARN (Pedregosa et al. 2011) implementation of logistic regression with balanced classes.

A2 Convolutional neural networks

CNNs are a biologically inspired prediction model for prediction with image inputs. The input image is convolved with a number of filters to produce output images called feature maps. These feature maps can then be convolved again with other filters on subsequent layers, producing a network of convolutions. The whole network is differentiable with respect to the values of the filters and the filters can be learned using gradient-based optimization methods. The final layer of the network is logistic regression, with the convolved outputs as input features. For more detail, see LeCun et al. (1998, subsection II.A). We use KERAS (Chollet et al. 2015) to implement our CNN, accounting for class imbalance by reweighting the classes.

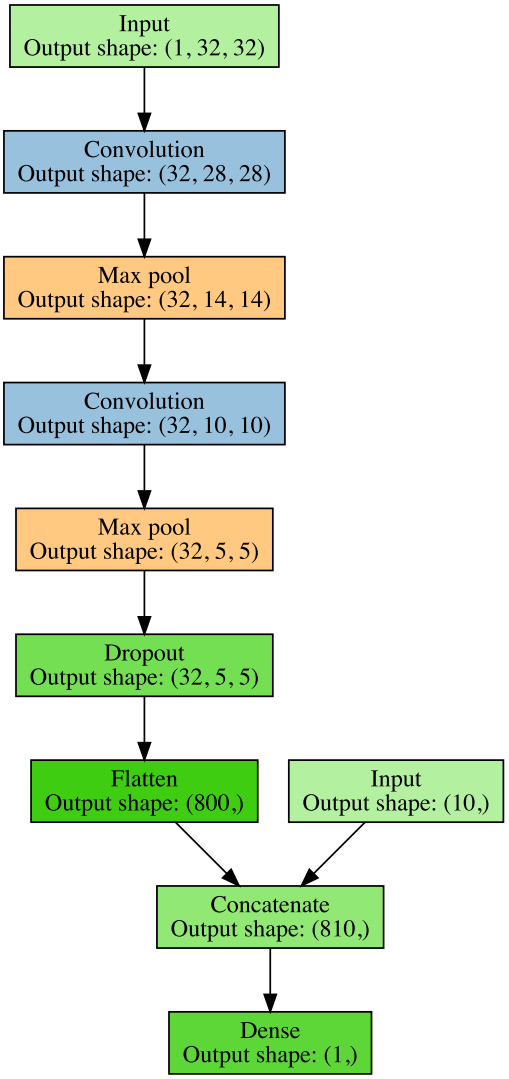


Figure A1. Architecture of our CNN. Parenthesized numbers indicate the size of output layers as a tuple (width, height, and depth). The concatenate layer flattens the output of the previous layer and adds the 10 features derived from the candidate host in SWIRE, i.e. the flux ratios, stellarity indices, and distance. The dropout layer randomly sets 25 per cent of its inputs to zero during training to prevent overfitting. Diagram based on https://github.com/dnouri/nolearn.

CNNs have recently produced good results on large image-based data sets in astronomy (e.g. Dieleman et al. 2015; Lukic et al. 2018). We employ only a simple CNN model in this paper as a proof of concept that CNNs may be used for class probability prediction on radio images. The model architecture we use is shown in Fig. A1.

A3 Random forests

Random forests are an ensemble of decision trees (Breiman 2001). They consider multiple subsamples of the training set, where each subsample is sampled with replacement from the training set. For each subsample, a decision tree classifier is constructed by repeatedly making axis-parallel splits based on individual features. In a random forest, the split decision is taken based on a random subset of features. To classify a new data point, the random forest takes the weighted average of all classifications produced by each decision tree. We used the SCIKIT-LEARN (Pedregosa et al. 2011) implementation of random forests with 10 trees, the information entropy split criterion, a minimum leaf size of 45 and balanced classes.

APPENDIX B: ACCURACY TABLES

This section contains tables of accuracy for our method applied to CDFS and ELAIS-S1. In Tables B1 and B2, we list the balanced accuracies of classifiers on the cross-identification task for CDFS and ELAIS-S1, respectively, averaged over each set of training quadrants. In Tables B3 and B4, we list the balanced accuracies of classifiers on the cross-identification task for CDFS and ELAIS-S1 respectively, averaged over each set of training quadrants.

Table B1. Balanced accuracies for different binary classification models trained and tested on SWIRE objects in CDFS. The 'Labeller' column states what set of training labels were used to train the classifier, and the 'Classifier' column states what classification model was used. 'CNN' is a convolutional neural network, 'LR' is logistic regression, and 'RF' is random forests. Accuracies are evaluated against the expert label set derived from Norris et al. (2006). The standard deviation of balanced accuracies evaluated across the four quadrants of CDFS (Fig. 8) is also shown. The 'compact' set refers to SWIRE objects within 1 arcmin of a compact radio component, the 'resolved' set refers to SWIRE objects within 1 arcmin of a resolved radio component, and 'all' is the union of these sets.

Labeller	Classifier	Mean 'compact' accuracy (per cent)	Mean 'resolved' accuracy (per cent)	Mean 'all' accuracy (per cent)
Norris	LR	91.5 ± 1.0	93.2 ± 1.0	93.0 ± 1.2
	CNN	92.6 ± 0.7	91.2 ± 0.5	92.0 ± 0.6
	RF	96.7 ± 1.5	91.0 ± 4.5	96.0 ± 2.5
RGZ	LR	89.5 ± 0.8	90.5 ± 1.7	90.2 ± 0.8
	CNN	89.4 ± 0.6	89.6 ± 1.3	89.4 ± 0.5
	RF	94.5 ± 0.2	95.8 ± 0.4	94.7 ± 0.3

Table B2. Balanced accuracies for different binary classification models trained on SWIRE objects in CDFS and tested on SWIRE objects in ELAIS-S1. Columns and abbreviations are as in Table B1. Accuracies are evaluated against the expert label set derived from Middelberg et al. (2008). The standard deviations of balanced accuracies of models trained on the four subsets of CDFS (Fig. 8) are also shown.

Labeller	Classifier	Mean 'compact' accuracy (per cent)	Mean 'resolved' accuracy (per cent)	Mean 'all' accuracy (per cent)
Norris	LR	94.6 ± 0.4	93.3 ± 2.0	95.3 ± 0.1
	CNN	94.8 ± 0.2	92.8 ± 0.5	94.4 ± 0.2
	RF	85.9 ± 3.8	70.0 ± 2.8	86.6 ± 3.2
RGZ	LR	91.8 ± 0.3	91.9 ± 0.5	92.0 ± 0.2
	CNN	90.1 ± 0.3	91.1 ± 0.9	90.2 ± 0.3
	RF	95.1 ± 0.1	95.2 ± 0.0	95.2 ± 0.3

Table B3. Cross-identification accuracies for different classification models on CDFS. The 'Labeller' column states what set of training labels were used to train the method, and the 'Classifier' column states what classification model was used. 'CNN' is a convolutional neural network, 'LR' is logistic regression, 'RF' is random forests, and 'Labels' is the accuracy of the label set itself. 'Perfect' indicates that the true labels of the test set were used and hence represents an upper bound on cross-identification accuracy with our method. 'NN' is a nearest-neighbours approach. Accuracies are evaluated against the expert label set, so 'Norris' labels are 100 per cent accurate by definition. The standard deviation of accuracies evaluated across the four quadrants of CDFS (Fig. 8) is also shown.

Labeller	Classifier	Mean 'compact' accuracy (per cent)	Mean 'resolved' accuracy (per cent)	Mean 'all' accuracy (per cent)
–	NN	97.2 ± 1.7	75.7 ± 7.9	93.4 ± 0.8
–	Random	97.9 ± 2.2	22.3 ± 9.2	83.2 ± 4.7
Norris	Labels	100.0 ± 0.0	100.0 ± 0.0	100.0 ± 0.0
	Perfect	97.9 ± 2.2	99.0 ± 1.8	98.1 ± 1.7
	LR	97.3 ± 0.5	76.0 ± 3.2	93.7 ± 1.8
	CNN	96.6 ± 0.9	74.3 ± 12.3	93.5 ± 0.5
	RF	96.1 ± 1.4	75.8 ± 6.7	93.8 ± 2.0
RGZ	Labels	53.1 ± 8.5	56.7 ± 5.9	54.4 ± 5.9
	LR	97.3 ± 1.9	74.5 ± 5.1	93.6 ± 1.7
	CNN	85.4 ± 2.6	68.1 ± 9.2	92.4 ± 1.1
	RF	97.5 ± 0.9	74.3 ± 7.9	93.7 ± 1.5

Table B4. Cross-identification accuracies for different classification models on ELAIS-S1. Columns and abbreviations are as in Table B3. Accuracies are evaluated against the expert label set derived from Middelberg et al. (2008) cross-identifications. The standard deviation of accuracies evaluated across models trained on the four quadrants of CDFS (Fig. 8) is also shown.

Labeller	Classifier	Mean 'compact' accuracy (per cent)	Mean 'resolved' accuracy (per cent)	Mean 'all' accuracy (per cent)
–	NN	95.5 ± 0.0	92.8 ± 0.0	95.5 ± 0.0
–	Random	61.9 ± 1.1	26.6 ± 2.1	61.9 ± 1.1
Middelberg	Perfect	99.6 ± 0.0	99.8 ± 0.0	99.6 ± 0.0
Norris	LR	89.0 ± 1.1	89.7 ± 1.8	94.4 ± 0.9
	CNN	89.7 ± 0.3	89.4 ± 1.4	94.3 ± 0.7
	RF	83.8 ± 5.6	82.3 ± 4.1	90.6 ± 2.1
RGZ	LR	90.5 ± 1.0	92.7 ± 0.2	95.9 ± 0.1
	CNN	84.6 ± 0.6	84.6 ± 0.6	91.8 ± 0.3
	RF	91.3 ± 1.0	90.3 ± 2.4	94.7 ± 1.2

APPENDIX C: SWIRE OBJECT SCORES

This section contains scores predicted by our binary classifiers for each SWIRE object within 1 arcmin of a radio component in CDFS and ELAIS-S1. Scores for SWIRE CDFS objects are shown in Table C1 (available online) and scores for SWIRE ELAIS-S1 are shown in Table C2 (available online). For CDFS, the score for an object in a quadrant is predicted by binary classifiers trained on all other quadrants. For ELAIS-S1, we show the scores predicted by binary classifiers trained on each CDFS quadrant. Note that these scores have *not* been weighted by Gaussians.

The columns of the score tables are defined as follows:

(i) *SWIRE* – SWIRE designation for candidate host galaxy.
(ii) *RA* – Right ascension (J2000).
(iii) *Dec* – Declination (J2000).
(iv) *Expert host* – Whether the candidate host galaxy is a host galaxy according to Norris et al. (2006) or Middelberg et al. (2008) cross-identifications of CDFS and ELAIS-S1, respectively.
(v) *RGZ host* – Whether the candidate host galaxy is a host galaxy according to RGZ cross-identifications (Wong et al. in preparation). This is always 'no' for ELAIS-S1 objects.
(vi) *C*(*L* / *D*) – Score assigned by binary classifier *C* trained on label set *L* of *D* candidate host galaxies. *C* may be 'CNN', 'LR' or 'RF' for CNN, logistic regression or random forests, respectively. *L* may be 'Norris' or 'RGZ' for expert and Radio Galaxy Zoo labels, respectively. *D* may be 'all', 'compact' or 'resolved' for each respective subset defined in Section 3.6.

APPENDIX D: ATLAS COMPONENT CROSS-IDENTIFICATIONS

This section contains cross-identifications predicted by our method for each ATLAS radio component in CDFS and ELAIS-S1. Cross-identifications for ATLAS CDFS components are shown in Table D1 (available online) and cross-identifications for ATLAS ELAIS-S1 are shown in Table D2 (available online). For CDFS, the cross-identification for a component in a quadrant is predicted using our method with binary classifiers trained on all other quadrants. For ELAIS-S1, we show the cross-identifications predicted by our method using binary classifiers trained on each CDFS quadrant. For CDFS, we also show the RGZ consensus, which is a proxy for the difficulty of cross-identifying a component (Wong et al. in preparation).

The columns of the cross-identification tables are defined as follows:

(i) *ATLAS* – ATLAS designation of radio component.
(ii) *RA* – Right ascension of radio component (J2000).
(iii) *Dec* – Declination of radio component (J2000).
(iv) *CID* – RGZ component ID.
(v) *Zooniverse ID* – RGZ Zooniverse ID.
(vi) *Norris/Middelberg* – Designation of SWIRE cross-identification from Norris et al. (2006) or Middelberg et al. (2008) for CDFS and ELAIS-S1 respectively.
(vii) *Norris/Middelberg RA* – Right ascension of SWIRE cross-identification from Norris et al. (2006) or Middelberg et al. (2008) for CDFS and ELAIS-S1 respectively.
(viii) *Norris/Middelberg Dec* – Declination of SWIRE cross-identification from Norris et al. (2006) or Middelberg et al. (2008) for CDFS and ELAIS-S1 respectively.
(ix) *RGZ* – Designation of SWIRE cross-identification from RGZ (Wong et al. in preparation).
(x) *RGZ RA* – Right ascension (J2000) of SWIRE cross-identification from RGZ (Wong et al. in preparation).
(xi) *RGZ Dec* – Declination (J2000) of SWIRE cross-identification from RGZ (Wong et al. in preparation).
(xii) *RGZ radio consensus* – percentage agreement of RGZ volunteers on the radio component configuration.
(xiii) *RGZ IR consensus* – percentage agreement of RGZ volunteers on the host galaxy of this radio component.
(xiv) *C*(*L* / *D*) – Designation of SWIRE cross-identification made by our method using classification model *C* trained on label set *L* of *D* candidate host galaxies. *C* may be 'CNN', 'LR' or 'RF' for CNN, logistic regression or random forests respectively. *L* may be 'Norris' or 'RGZ' for expert and RGZ labels respectively. *D* may be 'All', 'Compact' or 'Resolved' for each respective subset defined in Section 3.6.

(xv) *C*(*L* / *D*) *RA* – Right ascension (J2000) of SWIRE cross-identification made by our method using classification model *C* trained on label set *L* of *D* candidate host galaxies. *C*, *L* and *D* are defined as for designation.

(xvi) *C*(*L* / *D*) *Dec* – Declination (J2000) of SWIRE cross-identification made by our method using classification model *C* trained on label set *L* of *D* candidate host galaxies. *C*, *L* and *D* are defined as for designation.

APPENDIX E: CROSS-IDENTIFICATION FIGURES

This section contains figures of cross-identifications of each ATLAS radio component in CDFS and ELAIS-S1.

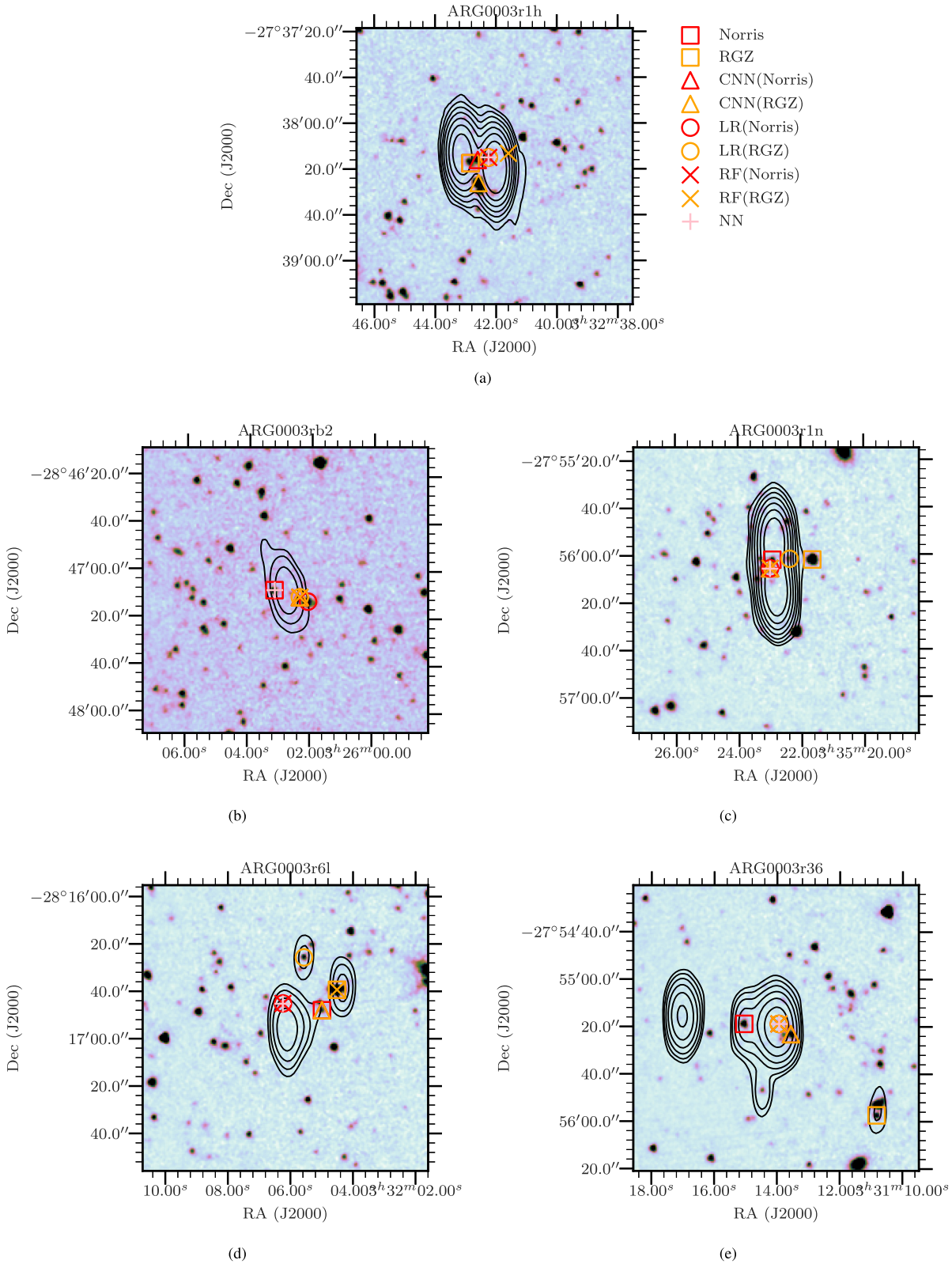


Figure E1. Examples of resolved sources with high disagreement between cross-identifiers. The contours show ATLAS radio data and start at 4σ, increasing geometrically by a factor of 2. The background image is the 3.6 μm SWIRE image. Binary classifier model/training set combinations are denoted $C(S)$ where C is the binary classifier model and S is the training set. 'LR' is logistic regression, 'CNN' is convolutional neural networks, and 'RF' is random forests. 'Norris' refers to the expert labels and 'RGZ' refers to the Radio Galaxy Zoo labels. The cross-identification made by nearest-neighbours is shown by 'NN'. The complete set of figures for 469 examples is available in the supplementary information.

This paper has been typeset from a TEX/LATEX file prepared by the author.